Regulatory-Compliant Data Management

Radu Sion*
Department of Computer Science
Stony Brook University
sion@cs.stonybrook.edu

Marianne Winslett†
Department of Computer Science
University of Illinois at Urbana Champaign
winslett@cs.uiuc.edu

1. OVERVIEW

Digital societies and markets increasingly mandate consistent procedures for the access, processing and storage of information. In the United States alone, over 10,000 such regulations can be found in financial, life sciences, health - care and government sectors, including the Gramm - Leach - Bliley Act, Health Insurance Portability and Accountability Act, and Sarbanes - Oxley Act. A recurrent theme in these regulations is the need for regulatory - compliant data management as an underpinning to ensure data confidentiality, access integrity and authentication; provide audit trails, guaranteed deletion, and data migration; and deliver Write Once Read Many (WORM) assurances, essential for enforcing long - term data retention and life - cycle policies.

While each regulation has its own unique characteristics, certain assurance features are broadly mandated:

Guaranteed Data Retention. The goal of compliant data management is to support WORM semantics: once written, data cannot be undetectably altered or deleted before the end of their regulation - mandated life span, even with physical access to its host.

Quick Lookup and Queries. In light of the massive amounts of data subject to compliance regulations, the regulatory requirement for quick data retrieval can only be met by accessing the data through indexing structures. Such indexes must be efficient enough to support a target throughput, and must be secured against insiders who wish to remove or alter compromising information before the end of its mandated lifespan.

Secure Deletion. Once data has reached the end of its lifespan, it can (and in some cases must) be deleted. Deleted records should not be recoverable even with unrestricted access to the underlying medium; moreover, after data is deleted, no hints of its existence should remain on the server, even in the indexes. We use the term *secure deletion* to describe this combination of features.

Compliant Data Migration. Retention periods are measured in years. For example, national intelligence information, educational records, and certain health records have retention periods of over 20 years. To address this requirement, compliant data management needs *data migration* mechanisms that allow information to be transferred from obsolete to new storage media while preserving its associated security guarantees.

Litigation Holds. Even if a data record has reached the end of its lifespan, it should remain fully accessible if it is the subject of current litigation.

Data Confidentiality. Only authorized parties should have access to compliance data. To meet this requirement, access should be restricted even if the storage media are stolen, and access to meta - data such as indexes should also be limited.

In addition to these features, a common thread running through many of these regulations is the perception of powerful insiders as the primary adversary. These adversaries have superuser powers coupled with full access to the storage system hardware. This corresponds to the perception that much recent corporate malfeasance has been at the behest of CEOs and CFOs, who also have the power to order the destruction or alteration of incriminating records. Since the visible alteration or destruction of records is tantamount to an admission of guilt in the context of litigation, a successful adversary must perform their misdeeds *undetectably*.

Unfortunately, current compliance data management mechanisms are vulnerable to faulty behavior or insiders with incentives to alter stored data because they rely on simple enforcement primitives such as software and/or device - hosted on/off switches, ill-suited to their target threat model. In practice, these first - generation mechanisms allow an insider using off - the - shelf resources to replicate illicitly modified data onto seemingly - identical units without detection.

More generally, the design of compliance data management is extremely challenging due to the conflict between security, cost - effectiveness, and efficiency. For example, the requirement to find requested information quickly means in practice data must be indexed. But trustworthy indexing of compliance data is a challenging problem, as it is easy to tamper with traditional indexes stored on WORM. Further, trustworthy indexes will make it very hard to delete all traces of documents that are past their retention periods, as required by certain regulations. Yet another complicating

*Radu Sion is supported partly by the NSF through awards CT CNS-0627554, CT CNS-0716608 and CRI CNS 0708025. Sion also wishes to thank Motorola Labs, IBM Research, CEWIT, and the Stony Brook VP for Research.
†Marianne Winslett is supported by the NSF under grants IIS-0331707, 0331690, CNS-0325951, and CNS-0524695.

VLDB '07, September 23-28, 2007, Vienna, Austria.

factor is the decades - long retention periods required by many regulations; it is unrealistic to expect data to reside on the same device for so long.

In this tutorial, we will discuss achieving strongly compliant data management in realistic adversarial settings. Specifically, we will explore designs for compliant data management systems that offer guaranteed document retention and deletion, quick lookup, and compliant migration, together with support for litigation holds and several key aspects of data confidentiality. Moreover, we will discuss the benefits of the recent advent of tamper - resistant, general - purpose trustworthy hardware which opens the door to fundamentally new assurance paradigms, e.g., by deploying this new hardware running certified code at the data management server. As heat - dissipation concerns greatly limit the performance of tamper - resistant processors, our goal is to investigate and evaluate software architectures for leveraging a secure processor in the server stack with minimal impact on cost and efficiency.

However, trusted hardware devices are not a panacea. Their practical limitations pose a set of significant challenges in achieving sound regulatory - compliance assurances. Specifically, heat dissipation concerns under tamper - resistant requirements limit the maximum allowable spatial gate - density. As a result, general - purpose secure coprocessors (SCPUs) are often significantly constrained in both computation ability and memory capacity, being up to one order of magnitude slower than host CPUs.

Such constraints mandate careful consideration in achieving efficient protocols. Direct implementations of the full processing logic *inside* the SCPU are bound to fail in practice due to lack of performance. The server's main CPUs will remain starkly under - utilized and the entire cost - proposition of having fast untrusted main CPUs and expensive slower secured CPUs will be defeated. Efficient protocols need to access the secure hardware sparsely, asynchronously from the main data flow.

We will explore these challenges and show how to leverage this new paradigm to achieve strong regulatory compliance for storage systems in realistic adversarial settings. Specifically, we will discuss achieving secure designs offering guaranteed record retention and deletion, quick lookup, and compliant migration, together with support for litigation holds and several key aspects of data confidentiality.

Recent compliance regulations are intended to foster and restore humans trust in digital information records and, more broadly, in our businesses, hospitals, and educational enterprises. As increasing amounts of information are created and live digitally, compliance data management will be a vital tool in restoring this trust and ferreting out corruption and data abuse at all levels of society.

2. STRUCTURAL OUTLINE

The tutorial will discuss three distinct compliance assurances, namely: (1) record - level Write - Once Read - Many (WORM) assurances, (2) trustworthy indexing, and (3) secure deletion. Specifically, in (1) we discuss existing tape-, optical-, and disk - based WORM mechanisms. We then explore the main drawbacks and vulnerabilities of such solutions and discuss achieving secure designs . In (2) we analyze existing mechanisms for secure indexing for various media and their vulnerabilities in the considered adversarial model . We then discuss indexing mechanisms impervious to such

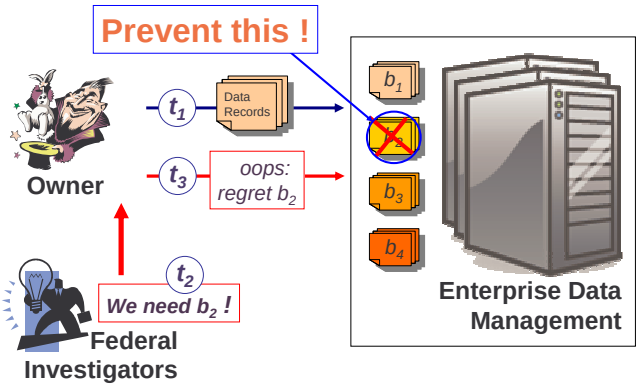

Figure 1: WORM prevents history "re-writing".

attacks, specifically Generalized Hash Trees (GHT) , supporting exact - match lookups of records based on attribute values, most suitable for use with structured data. We will then explore also unstructured domains and discuss thrustworthy keyword search , while surveying main results in the established area of thrustworthy indexing in outsourced data , including authenticated dictionaries , and query correctness mechanisms . We will discuss indexing mechanisms with secure hardware awareness and explore how such mechanisms can interact with the underlying WORM layer, discussed in (1). In (3), we will analyze the fact that, upon data record disposal (as mandated by numerous regulations) just erasing records from WORM is insufficient, as their contents (or artifacts thereof) may be recoverable from indexes. We will then explore the main challenges associated with this task as well as a set of emerging solutions , including logical deletion methods, history independent data structures and trusted hardware - aware mechanisms. Finally the tutorial also includes a brief general - audience introduction to main data security primitives.

3. BIOGRAPHY OF SPEAKERS

Radu Sion is an assistant professor of Computer Sciences in Stony Brook University and the director of the Network Security and Applied Cryptography Laboratory. His research focuses on data security and information assurance mechanisms. Collaborators and funding partners include Motorola Labs, IBM Research, the Center of Excellence in Wireless and Information Technology CEWIT, the Stony Brook Office for the Vice - President for Research and the National Science Foundation.

Marianne Winslett received her PhD in Computer Science from Stanford University in 1987. She has been an assistant, associate, full, and adjunct professor in the Department of Computer Science at the University of Illinois. Her research interests are in databases and related areas, especially security in open systems and parallel I/O for high - performance scientific computation. She received a Presidential Young Investigator Award from the National Science Foundation in 1989 and Xerox Awards for Faculty Research in 1990 and 1997. She is currently on the editorial board of ACM Transactions on Database Systems and is a former editor for IEEE Transactions on Knowledge and Data Engineering and the vice - chair of ACM SIGMOD.